SimVerity: When Does Simulated Agent Success Survive Physical Deployment?

Zhonghao Zhan[1], **Krinos Li**[1], **Yefan Zhang**[2], **Hamed Haddadi**[1]

[1]Imperial College London
[2]Independent Researcher

Abstract

Simulated evaluation is widely used to benchmark AI agents, yet how much evidence a simulated pass provides about physical deployment has not been systematically quantified. We present SimVerity, a verdict-transfer assurance framework: it replays matched scenarios on target smart home deployments and cross-validates agent execution against independently qualified physical witnesses. Our evaluation highlights that deployment success is a real-world process, not a static property in simulation: completion, reported state, observable effect, and settled outcome diverged within the same execution. Although an advanced simulator cleared all 240 light trials, a camera caught 42 sub-second failures invisible to settled-state checks. False clearance was predictable: a risk profile learned from measured trials and locked before evaluation predicted failures on a path it never physically measured, beating a property-blind baseline in all eleven held-out sessions across two cohorts. Agent auditability was also measurable: switching one agent loop's model–client/serving configuration raised its scenario-matching share from 52–88% to 100%. Finally, a second qualified simulator added no independent cross-check: it never disagreed on any overlapping case, and only physical measurement exposed their shared blind spots. SimVerity turns verdict transfer into an explicit decision: clear, abstain, or escalate before deployment.

Introduction

As autonomous AI agents transition into physical domains, they now command real homes. Google Home's agent executes blinds and lights as one spoken command, and lets camera scene understanding trigger automations across the home (Google Home 2026). Before such an agent ships, its green check comes from simulation: smart home benchmarks provide executable devices and state-based checks (Seo et al. 2025; Rivkin et al. 2023; Li et al. 2025a, 2026; Gu et al. 2026), tool agents receive sandbox verdicts (Yao et al. 2024), and world models score rollouts (Quevedo et al. 2025; Li et al. 2025b). A fundamental question remains unaddressed: does a simulator's pass justify physical deployment? When it does not, the false clearance, an approval whose property then fails, is an oversight failure, not benchmark noise (Kapoor et al. 2024; Xue et al. 2025; Ruan et al. 2024).

The root cause of false clearance is an abstraction mismatch: simulation sees success as a static property, while physical deployment unfolds as a process. The stakes are ordinary: in a basic agentic smart-home routine, the camera sees the car leave and the house shuts itself down, lights off, the space heater's smart plug off. Where stakes are obvious, vendors hard-code defenses: smart locks verify their own bolt, relock on 30-second timers, and are polled every 30 seconds by Home Assistant (Schlage 2026; U-tec 2026; Home Assistant 2026). However, a heater on a plug has only its reported state and whatever happens to be watching the room. Because false clearance cannot be rehearsed on real burners, we measure the same anatomy where failure is harmless: an agent issues `turn_off` and checks after 100 ms, the declared *read boundary*. The simulator passed all five held-out `off` trials and software state usually agreed, yet a qualified camera still saw the light on in 5/5; at the settled boundary, all 30 trials passed, including those 5. Completion, reported state, observable effect, and settled effect had been collapsed into one word: success. In SimuHome itself, a 3-second evaluation delay moves one agent's light-transition success from 44% to 62% (Seo et al. 2025, App. K). Success depends on when and where it is observed.

Figure 1: Instrumented Smart Home deployment.

Existing verification paradigms fall short of this property–verdict discrepancy. Simulation methodology insists a model is valid only for a stated purpose (Sargent 2013); conformance theory for cyber-physical systems proves which properties carry from model to system, but under dynamics assumptions that semantic, stochastic agent traces do not satisfy (Abbas, Mittelmann, and Fainekos 2014; Roehm et al. 2019); robotics and driving studies ask whether simulations predict reality, not whether one specific pass survives (Kadian et al. 2020; Li et al. 2024; Fremont et al. 2020). Closest in spirit, an admissibility ladder names "verdict-transfer-validated" as its highest level but leaves it unimplemented for lack of real-world anchors (Oefinger et al. 2026). None

supplies a reusable audit at property-verdict granularity.

To bridge this gap, we introduce SimVerity, an engine-agnostic *verdict-transfer assurance layer* outside the agent harness that quantifies whether a simulator's verdict survives on the target deployment. It reruns the declared scenarios on the deployment and grades each matched trace pair against the same declared property. It takes physical verdicts only from independently qualified witnesses (a camera must first prove it can tell the two outcomes apart); unqualified or missing evidence abstains rather than counting as agreement. From calibration runs alone it learns a frozen risk profile predicting which simulator passes will fail in deployment. The output is a card: risk, coverage, and abstain-or-escalate dispositions indexed by property $\times$ read boundary $\times$ path $\times$ deployment rung, the ladder Figure 2 draws from simulator to physical hardware. The layer never chooses actions or updates policy, and its witnesses are armed at actuation. We do not claim to have invented agent assurance or evaluator auditing; we identify verdict transfer as a missing assurance contract: paired, property-conditioned Verdict Fidelity and False Clearance Risk under hash-bound provenance.

We validate SimVerity on physical smart-home testbeds: three findings and a cross-check follow. First, deployed success is a process in the real world, not a property in simulation. Second, false clearance is predictable, and the simulator's own verdict is its worst predictor: a risk profile frozen before evaluation beat a property-blind baseline in all eleven held-out sessions across two cohorts, under natural ambient variation and curtain-controlled regimes alike; a deployment-local lookup stays stronger under stationary control. Third, auditability is measurable and configuration-conditioned: an unmodified production harness, Hermes (Nous Research 2026), kept every simulated trace matched to the declared scenarios; two custom agent loops, one ReAct-style, one planner-style, fell to 52–88%; and a registered model–client/serving intervention restored the same ReAct loop to 100%. Meanwhile, a second simulator changed coverage without ever disagreeing, leaving deployment anchors as the only independent check.

Methods

Problem Formulation

Let $x \sim \mathcal{D}$ be a legal scenario drawn from a declared evaluation distribution and ϕ a declared trace property. A *source* rung Q issues the clearance under audit, and a *target* rung T is where its survival is measured; after semantic alignment, $V_Q^\phi(x), V_T^\phi(x) \in \{\text{pass}, \text{fail}, \text{abstain}\}$. For binary eligible pairs, Verdict Fidelity is

$$\mathrm{VF}_\phi = \Pr_{x\sim\mathcal{D}}[V_Q^\phi(x) = V_T^\phi(x)], \tag{1}$$

and the deployment-critical error is False Clearance Risk,

$$\mathrm{FCR}_\phi = \Pr[V_T^\phi(x) = \text{fail} \mid V_Q^\phi(x) = \text{pass}]. \tag{2}$$

Abstain is a first-class outcome: it is assigned when a witness fails qualification, a trace stage is missing, or a capability is unsupported, and abstaining pairs are excluded from the binary estimands but tallied in the audit accounting, so missing evidence cannot silently become agreement. This paper instantiates three source–target pairings: direct simulator-to-physical calibration ($S \to R_{\text{phys}}$), software-rung-to-physical held-out prediction ($R_{\text{soft}} \to R_{\text{phys}}$), and retrospective replay of a second simulator against frozen physical anchors; the verdict-source column of Table 3 names Q for every evidence block. The interface is engine-agnostic: a learned world model, the judge class the admissibility ladder targets (Oefinger et al. 2026), is eligible as Q whenever its adapter exposes the manifest's actions, observations, and verdict-bearing property; present evidence uses public executable simulators, so that transfer remains an interface scope claim, not an empirical result. These estimands are conditional, not universal: a reported value answers one question, under the exact coordinates that produced it, from scenario distribution and read boundary to rungs and witness; change any coordinate and the value must be re-earned. This is operational validity at verdict granularity: the measured complement of a priori conformance guarantees, which are unavailable for semantic, stochastic agent traces.

System Overview

Figure 2 shows the audit design. A versioned manifest fixes the scenario and every coordinate of its execution, from read boundary and rungs to agent mode and split assignment. As scenarios execute on the source and target rungs, adapters collect traces without modifying either engine, alignment maps them onto semantic stages, witnesses qualify before testifying, and property monitors grade each matched pair; every audited cell exports a verdict-transfer card carrying VF_ϕ, FCR_ϕ, intervals, and a disposition: clear, abstain, or escalate. The contract is frozen before it is tested: every monitor, threshold, and risk profile is locked under SHA-256 hashes before any held-out ledger opens; ledgers are immutable, and the evidence chain terminates at explicit, falsifiable measurement assumptions.

Pipeline Components

Trace adapters and semantic alignment. Pairing is a bijection on trial identity: extra or missing trials must raise errors. Alignment maps each raw trace onto six semantic stages (request, source or ingestion report, agent read, dispatch, software feedback, observable effect); unavailable stages remain unavailable, and an ingestion time is never relabeled as the physical source time of a sensor event or effect.

Property monitors. A monitor is a total map from an aligned trace pair to $\{\text{pass}, \text{fail}, \text{abstain}\}$. The frozen set evaluates reported completion, reported read-after-write, observable effect after write, settled postcondition, effect order, and fanout completion; each declares its observation channel (reported state or qualified witness) and abstains, rather than passing, on a missing completion or boundary read, an invalid or missing witness, tied or missing effect times, or any unwitnessed fanout target. Sensor-to-effect scenarios reuse the read, observable, and settled monitors with a declared sensor cause; a cross-device cell asks whether an actuation leaves another device's reported channel quiet.

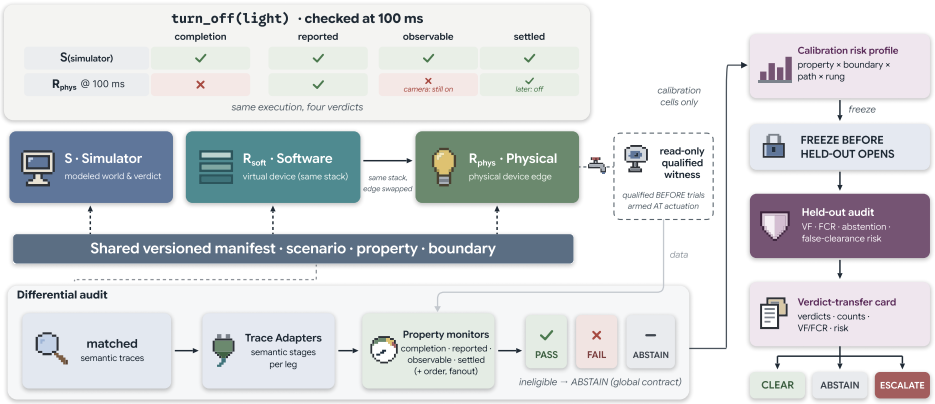

Figure 2: SimVerity audits a declared property across the simulator, R_{soft} (real stack, virtual edges), and R_{phys} (physical edges, prequalified read-only witness armed at actuation); a shared manifest aligns matched traces, and calibration-only cells freeze the risk profile before the held-out audit returns clear, abstain, or escalate.

Qualified witnesses and session validity. Instruments qualify before they testify. Let $\tilde{L}_{\text{on}}, \tilde{L}_{\text{off}}$ be median frame brightness over five settled frames per reference state; an optical witness qualifies only if $|\tilde{L}_{\text{on}} - \tilde{L}_{\text{off}}| \geq \max(10,\ 20\,\widehat{\text{MAD}}_{\text{pooled}})$, with midpoint and direction then frozen. Multi-target sessions further require, under randomized actuation, every absolute cross-response below half the target's own. A session is valid only with trace completeness $\geq 95\%$ and settled-control agreement; invalid initial states and missing witnesses abstain, and raw frames are reduced to scalar region statistics at capture and discarded. These thresholds qualify the instrument and never tune a boundary after outcomes. For example, the first pilot session during the experiment was invalidated by camera exposure drift and remains an audit record.

Calibration-only risk profiles. For each conditioning class c (property $\times$ stage or direction $\times$ path), the profile estimates deployment-failure risk from source-cleared calibration rows as the Jeffreys-smoothed rate $\hat{r}_c = (k_c + \frac{1}{2})/(n_c + 1)$, with n_c cleared pairs and k_c their deployment failures; held-out cells without support back off along a frozen conditioning ladder. Brier score (squared error of forecast risk) is primary because the output is a probability and calibration matters (Brier 1950); log loss, AUROC, coverage, and reliability plots are secondary. The design is intentionally simple: the question is whether declared structure carries across cells (property–boundary–path–rung combinations), not whether a learner can fit a testbed. The initial cohort's registered gate required beating a strong per-device latency lookup and a path-only profile on the held-out path–rung pairing; confirmatory cohorts registered path-only as primary criterion.

Audit Dispositions and Decision Logic

Every estimate ships with its full accounting: every clearance, block, and abstention is counted, session and path totals stay visible, and exact Clopper–Pearson intervals accompany each rate (Clopper and Pearson 1934). The accounting blocks two cheats: a judge that rarely clears and an audit that quietly discards hard-to-witness trials both show near-perfect agreement, and the carried counts expose them as collapsed clearance or abstention. The intervals are trial-level summaries, and the visible session and path counts delimit external inference: repeated trials on one deployment are not independent deployments. An assurance layer invites the question of who assures the assurer; the answer is that every element of the contract can fail visibly, and absent qualified evidence the audit abstains. Each audited cell exports a versioned verdict-transfer card that binds the run's identities and hashes to verdicts and counts, and closes with one of three dispositions: clear when evidence sustains the source's clearance, abstain when evidence fails qualification, escalate when cleared passes are observed to fail.

Experimental Design

Physical and Simulated Testbed

The audit runs on a multi-rung testbed: one primary simulator and one live deployment at two rungs. The simulator S is SimuHome, unmodified and driven only through an external adapter; the deployment runs Home Assistant on a live home, with the software rung R_{soft} routing commands through virtual device edges in the same stack and the physical rung R_{phys} terminating at real commodity devices (smart lights, battery contact sensors) under an independently qualified camera witness (Figure 1; see supplement for details). Four frozen paths carry the evidence: single-light effect (P1), multi-light order/fanout (P2), contact-sensor-to-lamp (P3), and cross-device observation (P5c, an append-only amendment within the read-after-write property). A path is reported as ineligible when its payload, source provenance, or witness cannot be qualified (per-path guardrails in the supplement).

Splits, Protocol, and Preregistered Gates

Calibration and held-out data are separated by session and cell, and repetitions between two device resets never cross the split: the recipe calibrates on the single- and multi-light paths at both rungs plus the R_{soft} sensor path and the cross-device cells, and holds out the sensor-to-effect path at R_{phys},

Table 1: Frozen predictors. Cohort 1's preregistered gate required lower held-out Brier than strong lookup and path-only; cohort 2 registered path-only as primary.

#	Predictor	Calibration information used
1	Simulator verdict	constant risk 0.01 per sim pass
2	Global FCR	one Jeffreys rate per property
3	Scalar latency	pooled effect latency by boundary
4	Strong lookup	quantiles per action/direction/device
5	Path-only	rates per path and boundary
6	SimVerity	property × stage/dir. × path

the pairing that carries the main prediction claim; unseen delays and actions serve as lesser feasibility levels, and an unseen simulator adapter is instantiated later by the append-only extension without reopening the physical program. The main campaign spans nine valid calibration sessions, 586 trials, and 1,070 eligible source-cleared pairs. The protocol starts at 10 repetitions per condition, finalizes counts from pre-freeze block variance, and requires at least 3 qualified physical sessions on the principal path. Session-aware paired bootstraps summarize ΔBrier with win/tie/loss, and exact intervals keep every path and session count visible. Adaptive test selection stays outside the frozen campaign, so it cannot contaminate the held-out gate.

Frozen Predictors

Table 1 lists the 6 frozen predictors as a ladder of structural granularity: from trusting the source's verdict outright, through property-blind rates and a per-device lookup, to the full property × stage/direction × path profile.

Agentic Replication and Second Simulator

Frozen plans isolate verdict transfer; live agents test whether the same evaluation contract remains applicable when actions are selected dynamically. The audited object is an agent's deployment clearance, not a policy learned by SimVerity. The study uses one local ReAct-style tool agent and one architecturally distinct planner–executor agent, frozen at temperature zero, with verdict metrics computed both conditional on matched actions and over the natural live-agent distribution; this is not a model leaderboard: the pre-registered auditability gate asks whether at least 80% of traces stay eligible and property-conditioned validity stays measurable under dynamic tool selection. After physical freeze, an append-only registration qualified a second public simulator, S5-HES (Siriweera et al. 2026), and paired both simulators' frozen-grid verdicts retrospectively against frozen P1 anchors (stored physical outcomes); an orchestration study compares custom ReAct, an unmodified production harness (Nous Research 2026), and the planner under one local configuration, and a frontier contrast repeats custom ReAct under another. Table 2 binds each measured or blocked cell to its exact model, route, and frozen count; the frontier intervention changes weights, provider endpoint, and serving stack together, so its unit is the full model–client/serving configuration. Stored pairs also compare single-simulator, consensus, and either-pass policies; physical evidence and the one-shot evaluator were never rerun.

Table 2: Executable agent configurations; rows index model–client/serving configurations, not model-only comparisons.

Config.	Model	Route / client	n	State
Physical deployment				
M2 ReAct	gpt-oss:20b	Ollama; native	32×2	invalid
M3 planner	deepseek-r1:32b	Ollama; native	32	valid
Simulation robustness (frozen 32-cell grid)				
ReAct	gpt-oss:20b	local `/v1`; custom	128	measured
Hermes v0.14.0	gpt-oss:20b	local `/v1`; Hermes	128	measured
Planner	gpt-oss:20b	local `/v1`; custom	128	measured
ReAct	grok-4.5	xAI API; custom	128	measured
ReAct / Hermes	gpt-5.6-sol	OpenAI API	0	blocked[a]
Hermes v0.14.0	grok-4.5	xAI API; Hermes	0	blocked[b]

Simulation rows: two trial orders, 12-turn budget; physical rows: frozen 32-trial manifests; every executed row enforces temperature 0. Blocked rows ($n = 0$): [a]provider rejects non-default temperature; [b]Hermes's custom-provider path fails silently and its native path exposes no enforceable identity.

Evidence and Calibration

Calibration: Pilot and Main Campaign

A precommitted pilot paired SimuHome, a single smart light, and a session-qualified camera witness (40 calibration and 30 held-out trials, 0 abstentions, predictor frozen first, all 22 registered checks independently reproduced): completion failed every trial, the settled control none, and the five held-out observable false clearances, all `off` trials at the unseen 0.1-s boundary, are the introduction's counterexample. The frozen profile's held-out Brier ran an order of magnitude below every pilot baseline: a feasibility result only, not satisfying the strong-baseline gate. The main campaign executed nine valid calibration sessions under the registered manifest: four at R_{soft} and five at R_{phys}, totaling 586 trials and 1,070 eligible simulator-pass pairs; exploratory pre-freeze sessions are excluded from every estimate.

Property-selective calibration structure. Across two devices and a grid extending to 0.05 s, the same executions yield four different conclusions (Figure 3(a,c) and Table 3). Settled effect is uniformly valid; observable failures are `off`-only through 0.25 s; reported state has its own sub-50 ms boundary; and completion is uniformly invalid at its declared read.

Ambient light decides who can testify. The motivating platform (Google Home) watches homes through cameras and a home's light changes all day, so physical sessions spanned the deployment's real lighting day, morning to dark. Ambient light then exercised the witness checks in both directions: two early calibration sessions were invalidated by their own pre-registered checks (optical separation between targets failed at dusk, on–off contrast at midday), and later

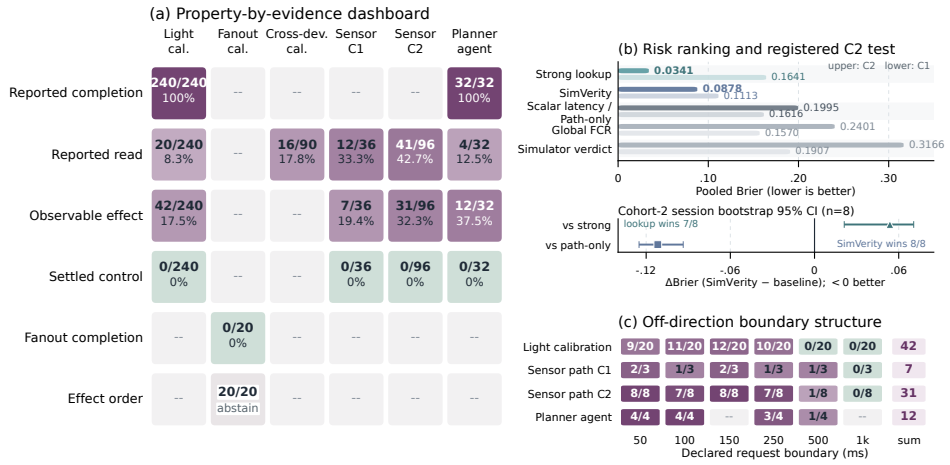

Figure 3: False clearances are property and boundary-selective. (a) Cards report false clearances/source-cleared eligible pairs. (b) Frozen Brier scores, cohort 2 upper and cohort 1 lower; scalar latency and path-only coincide (registered comparisons in Table 4). (c) Off-direction counts by boundary, with row sums at right.

agentic days refused or invalidated midday wash, dusk drift, full-dark infrared, and sunlit saturation. By the confirmatory cohort, ambient light had been promoted from nuisance to a curtain-controlled registered factor (next section).

Pre-registered replication blocks. Under a registered replication manifest, a television's reported state stayed stale after external power-on in every window (camera-based lower bounds near 160 s in the two cleanly anchored windows; supplement). Cold `turn_on` failed and warm retries succeeded intermittently; a success also repaired the stale reported state: acting is what made reading correct. A registered probe confirmed event multiplication (85 recorded transitions from 20 physical door-contact cycles); two suspected mechanisms, event merging and detector dead-time, were tested and ruled out.

Does the Audit Port? Site Replications and Scope

To test whether the audit and its discipline install beyond their birthplace, two further deployments, independently instrumented and preregistered on different device stacks, repeated the binary relay–lamp audit: Site B, a separately operated office with variable lighting conditions and a webcam (five qualified sessions, 120 trials), and Site C, a second home on a different vendor stack with an Apple TV, bedroom lamps, and a home security camera (two software-only sessions, 80 trials). Nothing is pooled across sites and the frozen predictor was never invoked: these sites are the paper's cross-stack evidence that the audit itself ships, its protocol and discipline running unchanged on deployments we did not build. Every completion, reported-state, and settled-postcondition pair passed: zero false clearances in their trials. The zeros are mechanism, not luck, and the mechanism is one simulators do not model: how an integration reports state. Both sites' relay integrations are confirmation-gated (reported state commits only on device acknowledgment), leaving no window to false-clear; Site C also measured the opposite, optimistic reporting that runs 50–500 ms ahead of the device. Instead of sites, semantics decides where reported state can lie. Cameras at both sites could not certify sub-second timing, so short-boundary optical cells abstained, exactly as the contract requires (per-site setups and trace bridges in the supplement). What the sites do not show is equally simple: they say nothing about how often such failures occur elsewhere, and they never ran the risk profile or the agents; the prediction evidence remains one home and one held-out pairing, 132 trials over eleven sessions. Everything beyond this operational scope rests on the one-shot, agentic, and cross-simulator tests reported next.

Results

Property-Selective Fidelity

The calibration card contains 1,090 rows: 1,070 binary pairs and 20 effect-order abstentions (camera cadence cannot resolve ordering). Within the same executions, completion can fail universally while settled effect never does, and both reported and observable verdicts vary by path. A scalar "task success" score would erase both the property failure and the coverage boundary.

Predicting False Clearance

After freezing six predictors, the held-out path–rung pairing ran once as an initial cohort: 36 source-cleared trials across three qualified R_{phys} sensor-to-effect sessions (a door contact dispatching a lamp). Verdicts on this path originate at R_{soft}, not the simulator, so the held-out test asks whether software-rung clearances survive the physical edge. Failures remained selective by direction and observation channel, not a uniform path defect (Table 3); Figure 3(b) and Table 4 carry the registered comparisons, and both prediction arrays are released for recomputation. All predictors covered 36/36; because the path–rung pairing was held out, path-only and SimVerity both predicted through frozen level-1 backoff, so the comparison tests whether the declared structure (Table 1) survives backoff. The profile won all three initial sessions against path-only and its pre-registered point-Brier

Table 3: Verdict-transfer ledger: false clearances/source-cleared eligible pairs; "–" = not evaluated. Cross-simulator rows reuse frozen P1 anchors. †At 0.05 s, S5-HES leaves 120/160 reported and observable replay pairs.

Block	Verdict source	Deployment pairing	Audit object	Completion	Reported	Observable	Settled
P1 calibration	SimuHome	live R_{phys}	plan (240)	240/240	20/240	42/240	0/240
P5c calibration	SimuHome	live R_{phys}	plan (90)	–	16/90	–	–
P3 held-out (one-shot)	R_{soft}	held-out R_{phys}	plan (36)	–	12/36	7/36	0/36
M3 live agent	SimuHome	live R_{phys}	M3 (32/32 match)	32/32	4/32	12/32	0/32
Cross-sim replay	SimuHome	frozen P1 anchors	plan (160)	160/160	20/160	30/160	0/160
Cross-sim replay	S5-HES	frozen P1 anchors	plan (160)	160/160	0/120†	21/120†	0/160

gate passed; against strong lookup it was point-better with resampling support spanning zero (supplement). A sign test cannot resolve three sessions (p=0.125), the thinness the confirmatory cohort was registered to close.

A preregistered confirmatory second cohort reran the same pairing: eight curtain-controlled sessions in one day (three ambient regimes), the same frozen predictor, the first cohort sealed, one evaluation written into a separate artifact. All eight attempts qualified, and the 96 fresh trials preserved the structure (observable FCR 31/96, reported 41/96, settled 0/96, 0 abstentions). The registered primary criterion was met decisively: 8/8 session wins against path-only, sign and exact Wilcoxon p=0.0039; across both cohorts the profile wins all eleven held-out sessions. The registered secondary comparison reversed: strong lookup beat the profile in seven of eight sessions. The two cohorts bracket a condition-dependent comparison: under natural ambient variation the structured profile was point-better with undecided support; under stationary controlled regimes the per-cell lookup memorized best. As a hypothesis, memorization pays under stationarity while declared structure is aimed at surviving condition shift. Six of eight sessions produced identical twelve-trial outcome vectors, reported openly; because the second protocol was registered with the first cohort's results known, the 11-session tally is descriptive, not a prospective joint test.

Auditability Under Live Agents

On frozen 32-trial manifests, the live-agent study replaces the plan with the M2 (ReAct) and M3 (planner–executor) physical configurations indexed in Table 2. Capture arms at the agent's final state-changing action; the witness capture time lands 0.393–0.739 s after the requested boundary in M3, neither the declared delay nor the capture time is hidden physical-effect time, and bounded capture under monotone dimming can only undercount short-boundary false clearances (supplement). Figures 3(c) and 4(a) report the boundary cells and auditability outcome; dry runs and gate-invalidated sessions are excluded from every estimate.

The pattern survives an architecturally distinct live agent. M3 passed all gates with zero abstention. Its observable failures remain `off`-only and boundary-graded, while reported state reproduces the sub-50 ms boundary (Figure 3(a,c)). The auditability gate therefore passes: the pattern remains measurable under dynamic tool selection; rates are not compared with the frozen-plan anchor because boundaries reference the agent's own final action. The planner's immediate return reads pre-update state (32/32 completion failures), whereas ReAct's closing turn lands after the update.

Table 4: Held-out prediction tests. Δ = SimVerity minus baseline (negative favors SimVerity); cohort-1 intervals are resampling support, cohort-2 bootstrap 95% CIs. Cohort 2 preregistered path-only as primary, strong lookup as secondary (no threshold).

	Cohort 1 (n=3)	Cohort 2 (n=8)
Simulator-verdict Brier	0.1907	0.3166
Path-only Brier	0.1616	0.1995
Strong-lookup Brier	0.1641	0.0341
SimVerity Brier	0.1113	0.0878
Wins vs path-only	3/3 (p=0.125)	8/8 (p=0.0039)
Δ vs path-only	$[-0.0940, -0.0077]$	$[-0.1251, -0.0936]$
Wins vs strong lookup	–	1/8
Δ vs strong lookup	$[-0.1724, +0.0203]$	$[+0.0212, +0.0707]$

Auditability is itself configuration-conditioned. M2 failed twice with identical failed-trial sets; Figure 4(a) shows the registered decomposition. With the shared local-model configuration fixed (Table 2), Figure 4(b) separates architectural labels from executable configurations: the production harness remains fully matched across both simulators while the custom loops cross the auditability gate in a simulator-dependent way (loss classes in the supplement). Switching the fixed custom ReAct loop and task grid to the registered frontier configuration restores every action match. This pre-registered full-configuration intervention falsifies ReAct itself as the stable cause, but cannot separate model weights from provider, endpoint, and serving behavior. Auditability is therefore a property of the executable configuration, not of a harness label.

Cross-Simulator Transfer and Clearance Policies

After physical freeze, a second simulator, S5-HES, qualified through its engine below the built-in LLM (Siriweera et al. 2026); its 100-ms tick structurally abstains at 0.05 s. The simulators never disagreed when mutually eligible, and Table 3 shows both inherit the same property-selective physical failures. Figure 4(c) exposes the policy result: strict consensus equals the stricter simulator, either-pass the permissive one; the apparent risk reduction is entirely forced abstention at the coarser clock, not detection by a second judge. On this grid the second green light supplied no independent opinion, and deployment anchors alone

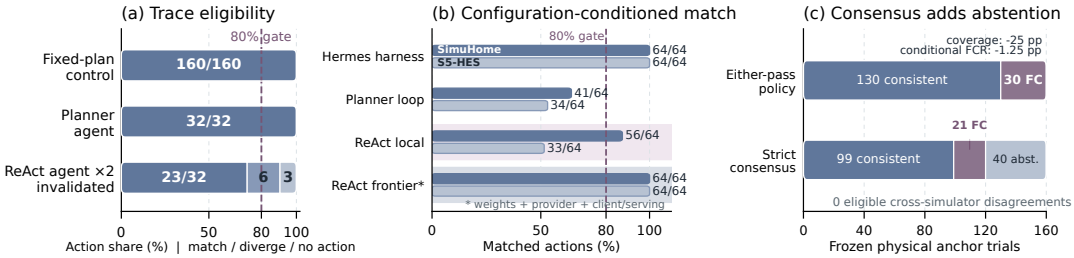

Figure 4: Auditability is configuration-conditioned; consensus changes coverage instead of detection. (a) Fixed-plan and planner traces fully match; both ReAct runs share a 23/6/3 matched/divergent/no-action split (no M2 rate). (b) Match rates pool two sessions per executable configuration, not model weights alone. (c) Zero eligible disagreements on 160 anchors: strict consensus lowers conditional FCR by 1.25 points by surrendering 25 coverage points.

calibrate risk; bridged onto two further preregistered deployments, the same frozen clearances incur zero reported or settled false clearances (supplement). Across these studies, no simulator-side substitute suffices: simulator agreement and stronger executable configurations each fail to replace deployment-grounded verdict assurance.

Related Work

Simulation validity and verdict preservation. Fit-for-purpose validity (Sargent 2013), CPS conformance (Abbas, Mittelmann, and Fainekos 2014; Roehm et al. 2019), and digital-twin trace alignment (Muñoz et al. 2024) formalize model trust; the L0–L4 ladder names verdict transfer as its highest rung (Oefinger et al. 2026), which SimVerity instantiates as per-property agreement, FCR, abstention, and held-out prediction.

Sim–real predictivity and physical testing. Sim-to-real studies test whether rankings predict reality (Kadian et al. 2020; Li et al. 2024; Truong et al. 2023), world-model evaluators automate rollout scores (Quevedo et al. 2025; Li et al. 2025b), and driving studies compare virtual and physical failures (Fremont et al. 2020; Stocco, Pulfer, and Tonella 2022; Haq et al. 2021); the per-scenario test-track comparison is the closest verdict-level precedent, and SimVerity makes VF/FCR reusable with probability quality tested.

Agent benchmarks and sandbox validity. Smart-home and tool agents receive saved-state or sandbox verdicts (Rivkin et al. 2023; Li et al. 2025a; Seo et al. 2025; Li et al. 2026; Gu et al. 2026; Yao et al. 2024), yet offline evaluation can mislead deployment (Kapoor et al. 2024; Xue et al. 2025). ToolEmu validates sim-fail to plausible-real-fail (Ruan et al. 2024); FCR asks whether sim-pass becomes observed deployment-fail. LLM personas and virtual ambient sensors synthesize smart-home traces (Jüttner and Buchmann 2026; Leng et al. 2026), enriching simulators whose verdicts still need auditing; a sim-to-real agenda and failure-injection benchmark (Liu et al. 2026; Zhou et al. 2026) motivate SimVerity's paired measurement layer.

Evaluation validity as an alignment problem. Alignment-track work audits the evaluation pipeline itself: contamination, steerability side effects, evaluator bias, and success–safety separation (Liang et al. 2026; Chang et al. 2026; Recchia et al. 2026; Lee et al. 2026). BenchGuard audits benchmark artifacts and DFAH measures replay determinism (Tu et al. 2026; Khatchadourian 2026); neither tests whether a simulator verdict survives witnessed deployment. Deployment simulation forecasts prevalence from prior traffic (Williams et al. 2026) and Composable Assurance propagates formal assertions through MLOps (Zhao 2026); SimVerity instead calibrates whether simulator verdicts remain deployment-admissible.

Limitations, Ethics, and Reproducibility

The primary predictor profile covers one held-out physical path–rung pairing, eleven sessions in two cohorts: it estimates risk, not certification, and the strong-lookup comparison splits by ambient condition, favoring the lookup under stationary control. The unpooled site cohorts test portability alone. Physical evidence has one valid planner configuration; invalid ReAct sessions mark a configuration-level auditability boundary. The model–client/serving intervention cannot isolate weights from serving, with frontier cells blocked (Table 2); ecosystems without first-class determinism resist frozen evaluation. Only whitelisted lights and isolated proxies were tested. Frames are reduced to scalars and discarded; no human data are collected. Versioned manifests, sanitized ledgers, hashes, and replay code regenerate all aggregates (abstentions/invalid sessions retained); frozen arrays recompute one-shot metrics; materials provided.

Conclusion

A green check from simulation is a promise about the physical world; we measured how often it holds. On a live home, the same passes split into four verdicts, and a camera kept catching lights that had not yet obeyed. The blind spots proved predictable: a risk profile frozen before testing anticipated them on a path it never physically measured, beating a property-blind baseline in all eleven held-out sessions across two cohorts. Auditability followed the executable configuration, and a second qualified simulator never disagreed with the first: only physical measurement exposed their shared blind spots. Ship the audit, not the answer: answers are deployment-local by construction.

References

Abbas, H.; Mittelmann, H.; and Fainekos, G. 2014. Formal property verification in a conformance testing framework. In *2014 Twelfth ACM/IEEE Conference on Formal Methods and Models for Codesign (MEMOCODE)*, 155–164. IEEE.

Brier, G. W. 1950. Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1): 1–3.

Chang, T.; Schnabel, T.; Swaminathan, A.; and Wiens, J. 2026. A course correction in steerability evaluation: revealing miscalibration and side effects in LLMs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37259–37267.

Clopper, C. J.; and Pearson, E. S. 1934. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4): 404–413.

Fremont, D. J.; Kim, E.; Pant, Y. V.; Seshia, S. A.; Acharya, A.; Bruso, X.; Wells, P.; Lemke, S.; Lu, Q.; and Mehta, S. 2020. Formal scenario-based testing of autonomous vehicles: From simulation to the real world. In *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*, 1–8. IEEE.

Google Home. 2026. Google Home Release Notes: May 27th. https://support.google.com/googlehome/thread/437171944. Accessed 2026-07-25.

Gu, Y.; Wang, H.; Zhang, S.; Hou, Y.; Xue, L.; Ming, W.; Liu, C.; Yu, F.; Li, K.; Chen, R.; et al. 2026. HomeFlow: A Data Flywheel for Smart Home Agent Training with Verifiable Simulation. *arXiv preprint arXiv:2606.01230*.

Haq, F. U.; Shin, D.; Nejati, S.; and Briand, L. 2021. Can offline testing of deep neural networks replace their online testing? a case study of automated driving systems. *Empirical Software Engineering*, 26(5): 90.

Home Assistant. 2026. Schlage integration. https://www.home-assistant.io/integrations/schlage. Accessed 2026-07-25.

Jüttner, V.; and Buchmann, E. 2026. Simulating the Resident: Generating Executable Smart Home Schedules via LLM Personas. *arXiv preprint arXiv:2607.08231*.

Kadian, A.; Truong, J.; Gokaslan, A.; Clegg, A.; Wijmans, E.; Lee, S.; Savva, M.; Chernova, S.; and Batra, D. 2020. Sim2real predictivity: Does evaluation in simulation predict real-world performance? *IEEE Robotics and Automation Letters*, 5(4): 6670–6677.

Kapoor, S.; Stroebl, B.; Siegel, Z. S.; Nadgir, N.; and Narayanan, A. 2024. Ai agents that matter. *arXiv preprint arXiv:2407.01502*.

Khatchadourian, R. 2026. Replayable Financial Agents: A Determinism-Faithfulness Assurance Harness for Tool-Using LLM Agents. *arXiv preprint arXiv:2601.15322*.

Lee, J.; Hahm, D.; Choi, J. S.; Knox, W. B.; and Lee, K. 2026. Mobilesafetybench: Evaluating safety of autonomous agents in mobile device control. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37565–37573.

Leng, Z.; Thukral, M.; Liu, Y.; Rajasekhar, H.; Hiremath, S. K.; He, J.; and Plötz, T. 2026. AgentSense: Virtual Sensor Data Generation Using LLM Agents in Simulated Home Environments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 1891–1899.

Li, K.; Zhang, S.; Wang, H.; Yu, F.; Sheng, Z.; Gu, Y.; Ming, W.; Xue, L.; Liu, C.; Hu, S.; et al. 2026. SMH-Bench: Benchmarking LLM Agents for Environment-Grounded Reasoning and Action in Smart Homes. *arXiv preprint arXiv:2606.01912*.

Li, S.; Guo, Y.; Yao, J.; Liu, Z.; and Wang, H. 2025a. HomeBench: Evaluating LLMs in smart homes with valid and invalid instructions across single and multiple devices. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 12230–12250.

Li, X.; Hsu, K.; Gu, J.; Pertsch, K.; Mees, O.; Walke, H. R.; Fu, C.; Lunawat, I.; Sieh, I.; Kirmani, S.; et al. 2024. Evaluating real-world robot manipulation policies in simulation.

Li, Y.; Zhu, Y.; Wen, J.; Shen, C.; and Xu, Y. 2025b. Worldeval: World model as real-world robot policies evaluator. *arXiv preprint arXiv:2505.19017*.

Liang, Z.; Yu, L.; Shiyu, Z.; Ye, Q.; and Hu, H. 2026. How much do large language model cheat on evaluation? benchmarking overestimation under the one-time-pad-based framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37636–37644.

Liu, X.; Chen, T.; Li, W.; Hu, X.; and Wei, H. 2026. The Sim-to-Real Gap of Foundation Model Agents: A Unified MDP Perspective.

Muñoz, P.; Wimmer, M.; Troya, J.; and Vallecillo, A. 2024. Measuring the fidelity of a physical and a digital twin using trace alignments. *IEEE Transactions on Software Engineering*, 50(12): 3122–3145.

Nous Research. 2026. Hermes Agent. Version 0.14.0.

Oefinger, C.; Schäfer, F. R.; Moller, K.; Piccinini, M.; and Betz, J. 2026. Validate the Dream Before You Trust Its Verdict: Admissibility for World-Model Simulators. *arXiv preprint arXiv:2607.07196*.

Quevedo, J.; Sharma, A. K.; Sun, Y.; Suryavanshi, V.; Liang, P.; and Yang, S. 2025. WorldGym: World Model as An Environment for Policy Evaluation. *arXiv preprint arXiv:2506.00613*.

Recchia, G.; Mangat, C. S.; Nyachhyon, J.; Sharma, M.; Canavan, C.; Epstein-Gross, D.; and Abdulbari, M. 2026. Confirmation bias: A challenge for scalable oversight. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 37877–37886.

Rivkin, D.; Hogan, F.; Feriani, A.; Konar, A.; Sigal, A.; Liu, S.; and Dudek, G. 2023. Sage: smart home agent with grounded execution. *arXiv preprint arXiv:2311.00772*.

Roehm, H.; Oehlerking, J.; Woehrle, M.; and Althoff, M. 2019. Model conformance for cyber-physical systems: A survey. *ACM Transactions on Cyber-Physical Systems*, 3(3): 1–26.

Ruan, Y.; Dong, H.; Wang, A.; Pitis, S.; Zhou, Y.; Ba, J.; Dubois, Y.; Maddison, C.; and Hashimoto, T. 2024. Identifying the risks of lm agents with an lm-emulated sandbox. In *International Conference on Learning Representations*, volume 2024, 27031–27098.

Sargent, R. G. 2013. Verification and validation of simulation models. *Journal of simulation*, 7(1): 12–24.

Schlage. 2026. How to use Schlage smart locks' auto-lock feature for a safer home. https://www.schlage.com/en/blog/product_updates/auto-lock-smart-locks.html. Accessed 2026-07-25.

Seo, G.; Yang, J.; Pyo, J.; Kim, N.; Lee, J.; and Jo, Y. 2025. SimuHome: A Temporal-and Environment-Aware Benchmark for Smart Home LLM Agents.

Siriweera, A.; Rangila, J.; Naruse, K.; Paik, I.; and Jayanada, I. 2026. S5-HES Agent: Society 5.0-driven Agentic Framework to Democratize Smart Home Environment Simulation. *arXiv preprint arXiv:2603.01554*.

Stocco, A.; Pulfer, B.; and Tonella, P. 2022. Mind the gap! a study on the transferability of virtual versus physical-world testing of autonomous driving systems. *IEEE Transactions on Software Engineering*, 49(4): 1928–1940.

Truong, J.; Rudolph, M.; Yokoyama, N. H.; Chernova, S.; Batra, D.; and Rai, A. 2023. Rethinking sim2real: Lower fidelity simulation leads to higher sim2real transfer in navigation. In *conference on Robot Learning*, 859–870. PMLR.

Tu, X.; Wang, T.; Huang, K.; Qu, Y.; Mostafavi, S.; et al. 2026. Benchguard: Who guards the benchmarks? automated auditing of llm agent benchmarks. *arXiv preprint arXiv:2604.24955*.

U-tec. 2026. How to set up auto lock for Ultraloq U-Bolt series and Bolt series. https://support.u-tec.com/hc/en-us/articles/360028480451. Accessed 2026-07-25.

Williams, M.; Sheahan, H.; Raymond, C.; Korbak, T.; Pan, D.; Yang, P.; Maksin, L.; Xie, N.; Guo, P.; Kivlichan, I.; et al. 2026. Predicting LLM Safety Before Release by Simulating Deployment. *arXiv preprint arXiv:2607.07184*.

Xue, T.; Qi, W.; Shi, T.; Song, C. H.; Gou, B.; Song, D.; Sun, H.; and Su, Y. 2025. An illusion of progress? assessing the current state of web agents.

Yao, S.; Shinn, N.; Razavi, P.; and Narasimhan, K. 2024. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains.

Zhao, X. 2026. Composable Assurance for AI Alignment: A Framework for Propagating Formal Safety Properties Through MLOps. 40(44): 38129–38136.

Zhou, X.; Yuan, A.; Luo, Z.; Ling, Z.; Pan, X.; Gao, Y.; Zhang, H.; Li, J.; Jiang, S.; Wang, P. Z.; et al. 2026. When Simulation Lies: A Sim-to-Real Benchmark and Domain-Randomized RL Recipe for Tool-Use Agents. *arXiv preprint arXiv:2605.11928*.